Combining Evidence for Automatic Extraction of Terms

Boris Dobrov and Natalia Loukachevitch

Research Computing Center of Lomonosov Moscow State University,
Leninskie Gory, 4/1, Moscow, Russia
`{dobroff,louk}@mail.cir.ru`

Abstract. The paper describes the method of extraction of two-word domain terms combining their features. The features are computed from three sources: the occurrence statistics in a domain-specific text collection, the statistics of global search engines, and a domain-specific thesaurus. The evaluation of the approach is based on the terminology of manually created thesauri. We show that the use of multiple features considerably improves the automatic extraction of domain-specific terms. We compare the quality of the proposed method in two different domains.

Keywords: term acquisition, thesaurus, Internet search, machine learning.

1 Introduction

Technical terms represent domain-specific concepts in domain documents. It means that terms are key elements to understanding of the conceptual structure of the domain. Therefore automatic extraction of domain-specific terms from texts is a subject of constant interest in automatic processing of documents. The special difficulty is the automatic extraction of multiword terms [1].

When working with large subject areas and text collections, even the best methods for extracting terminological collocations show significant drop in term percentage from 90% in the first hundred of the list, down to 60% in the third thousand. Thus, an important problem is development of new methods to improve the quality of phrase ordering to increase the percentage of terms at the top.

For many years researchers tried to find the best statistical feature for term extraction. Now machine learning methods allow for the combination of many features. Thus, in [2] the combination of different statistical characteristics of phrases, based on the Czech text collection, is used to extract different types of collocations (such as phrasal verbs or idioms). The authors used over 80 features and obtained 20% improvement compared with the best individual feature. But the authors of this paper indicate that efficiency of different features is very variable and depends on a collection, types of expressions and so on.

In this paper we describe an experiment to extract two-word terms based on a combination of three types of features: features based on a domain-specific text

collection, features obtained from an Internet search engine and features obtained from a domain-specific thesaurus. Working with a thesaurus we simulate the situation when a thesaurus partially exists. We want to study its potential to recognise new terms. The important point of our research is to study the stability of the term extraction model among different domains.

2 Description of Experiment: Data and Evaluation

We conduct our study in two domains. The first domain is the very broad domain of natural sciences and technologies. The second domain is domain of banking and bank regulation. For both domains we have Russian domain-specific thesauri developed manually, which we use as a basis for evaluation of term extraction methods. The current volume of Ontology on Natural Sciences is more than 140 thousand terms [3]. Banking thesaurus includes about 15 thousand terms.

Besides we have Russian domain-specific text collections used for development of these thesauri. From the text collections we have extracted single words and multiword expressions. Two-word expressions belong to two types of noun groups: Adjective+Noun and Noun+Noun_in_Genitive. The extracted expressions were ordered in descending order of their frequencies. Our experts worked with these term candidate lists paying more attention to expressions with high frequencies. However it was noted that the important terms could have medium or low frequencies because of the unbalance of the text collections. So the aim of our new term extraction method is to reorder the extracted expressions to provide the presence of more approved terms in the beginning of the candidate list. We experimented with five thousands of the most frequent two-word expressions from these candidate lists.

To evaluate the reordering quality of methods we use the measure of average precision adopted from information retrieval [4]. Average precision AvP in the task of extracting terminological expressions is calculated as follows. Suppose that in an ordered list of expressions there are k terms, and $\text{pos}(i)$ - position of the i -th term from the beginning of the list. Then the precision on the level of the i -th terminological expression PrecTerm_i in an ordered list is **PrecTerm(pos(i))**, that is the value of precision PrecTerm_i is calculated at the time of inclusion to the list of i -th term and is equal to the percentage of terms in the list from 1 to $\text{pos}(i)$ positions. Average precision for the given ordered list is equal to the average value of PrecTerm_i :

$$\text{AvP} = \frac{1}{k} \sum_i \text{PrecTerm}_i$$

Measure of average precision allows us to estimate the quality of term extraction with a single value because the more domain terms are located in the beginning of the list the more is the value of average precision.

3 Features for Term Candidate Reordering

For extracted phrases we compute features of three types. First, we use features calculated on the basis of the domain-specific text collection. In this experiment such features are **the frequency of the phrase**, **mutual information (MI)**, **MI3** (the same as MI but accounts the cube of the phrase frequency [5]) and **Insideness**. **Insideness** is calculated as the inverse ratio of the phrase frequency to maximal frequency of a three-word phrase comprising the given phrase.

We use not so many features from the domain-specific text collection because our main interest is to study other types of features based on Internet search and known terms described in a domain-specific thesaurus. We will consider these types of features in the following subsections.

3.1 Features Based on Internet Search

Internet-based features were obtained with xml-interface of Russian Search Engine Yandex on the basis of specially formulated queries. For our experiments we utilised so-called search snippets - short fragments of texts explaining search results. Example of one of Yandex snippets for phrase **internal debt** is as follows:

Internal debt is the part of a country's debts that is owed to creditors who are citizens of that country. It is a form of fiat creation of money, in which the government obtains cash not by printing it, but by borrowing it.

Use of Internet search for term extraction is important for the following reasons. First, the domain-specific text collection is always not sufficient because a lot of fairly significant terms of the domain may have relatively low frequencies in it. Involvement of the Internet helps us get additional information on such terms. Secondly, the use of information from the Internet allows us to find out if a given phrase is rigidly connected with the domain.

To calculate phrase features, 100 snippets from search results were utilised. Snippets from the same query were merged into one document and processed by the morphological processor. As a result, for each set of snippets lemmas (words in a dictionary form) were extracted and their frequencies of occurrence were calculated. For every query, we obtain a vector of lemmas with corresponding frequencies. Snippets are generated for the whole phrases and its constituent words. We denote S_{ab} - a vector of lemma frequencies derived from phrase snippets, S_a , S_b - vectors of lemmas from constituent word snippets. From such vectors, the following types of features were calculated.

The first group of features are scalar products of snippet vectors: $\langle S_{ab}, S_a \rangle$ (**Scalar1**), $\langle S_{ab}, S_b \rangle$ (**Scalar2**). Also we calculated scalar products of Boolean variants of snippet vectors (elements of vectors are from $\{0, 1\}$): $\langle Sb_{ab}, Sb_a \rangle$ (**Boolean1**), $\langle Sb_{ab}, Sb_b \rangle$ (**Boolean2**). These features compare overall distinctions in lexical contexts of a phrase and its components.

The frequency of a phrase in its own snippets (FreqBySnip). We supposed that if the value of this feature is significantly greater than 100 (sometimes

this feature reached 250-300 occurrences in 100 snippets), it means that there are many contexts, in which this phrase is explained in detail, is the theme of the fragment, and, most likely, this phrase denotes an important concept or a specific entity.

Features of semantically specific context ($SnipFreq_i$). Many domain-specific multiword terms have sense specificity, which can not be deduced from its components. This specificity of the term can be shown in additional relations to other terms not directly related to initial term components. To find such specificity we extract a single lemma that is very frequent in phrase snippets and absent (or rarely mentioned) in component snippets.

Let lemma L occur f_{ab} times in phrase snippets, f_a , f_b times occur in snippets of components. Then we calculate $SnipFreq_0$ feature as follows:

$$SnipFreq_0 = \max_L f_{ab-a-b} \log \frac{N - d_{Lcol}}{d_{Lcol}}$$

where

$$f_{ab-a-b} = \max(f_{ab} - f_a - f_b, 0),$$

dlcol is a lemma frequency in documents of a contrastive collection, **N** - is the number of documents in the contrastive collection. Factor $\log \left(\frac{N - d_{Lcol}}{d_{Lcol}} \right)$ is so called **idf-factor** known from information retrieval research [4], it helps to diminish influence of frequent general words. The contrastive collection is the collection of documents of Belorussian Internet distributed in the framework of Russian seminar on information retrieval (www.romip.ru/en/index.html).

$SnipFreq_1$ and $SnipFreq_2$ features are calculated in similar way excepting words in a window of 1 (2) words near every occurrence of ab phrase. These variants of $SnipFreq$ feature are intended to remove partial fragments of longer terms from consideration.

Number of definitional words in snippets (NearDefWords). This feature calculates overall frequency of so called definitional words in phrase snippets. These words (as type, class, define etc.) are often used in dictionary definitions. Therefore their presence in snippets can mean that a snippet contains a definition of this phrase or the phrase is used in definition of other term. **NearDefWords** feature is equal to the number of these definitional words that appeared immediately adjacent (left or right) with the original phrase in snippets.

Number of marker words in snippets (Markers). This feature denotes number of five-ten the most important words of the domain in snippets of the expression. For the natural science domain these words were as follows: mathematics, mathematical, physics, physical, chemistry, chemical, geology, geological, biology, biological.

Number of Internet page titles (SnipTitle). We calculated number of Internet page titles coinciding with a given phrase, because we supposed that use of the expression in the title of an Internet page stresses significance of the phrase.

3.2 Features Based on Terms of Domain-Specific Thesaurus

In many domains there are well-known terms and even information-retrieval thesauri. The third type of our features is based on the assumption that the known terms can help to predict unknown terms. For the experiments in two domains, we used the relevant thesauri. If a phrase was a thesaurus term, then it was excluded from the terminological basis for feature generation. We considered the following features obtained from a domain-specific thesaurus.

Synonym to Thesaurus Term (SynTerm). Documents of the domain can contain a lot of variants of the same term [6]. Therefore we can suppose that a phrase similar to a thesaurus term is also a term. Let a and b be components of phrase ab . We consider phrase cd as a synonym of ab phrase if every component word of phrase cd is either equal to a component word of ab either is a synonym of a component word of ab . The order of components in a phrase is unimportant.

Synonym to Non-Term (SynNotTerm). We also fix a feature of similarity to a phrase not included to the thesaurus.

Completeness of Description (Completeness). It is possible that component words a and/or b of phrase ab have been already described in the domain thesaurus as domain terms. Completeness feature is a sum of thesaurus relations of component terms in the thesaurus.

4 Results of Experiments

We experimented in two domains: the banking domain and the domain of natural sciences. In all experiments, 5 thousand most frequent two-word expressions extracted from the corresponding text collections were used. For these expressions all above-mentioned features were calculated. To obtain the best combination of features generating the best reordering, we used machine learning methods implemented in programming package RapidMiner (www.rapidminer.com). The quality of reordering was evaluated with AvP measure. The training set was three-quarters of the phrase list, the testing set was a quarter. As basic minimal levels of AvP we used the alphabet order and the decreasing frequency order.

We tested various machine learning methods from RapidMiner package. Every time logistic regression achieved maximal level of AvP. Therefore we took this method as a basic method of machine learning for our experiments in term extraction. The table 1 shows values of AvP for single features and results of logistic regression method.

SynTerm and **SynNotTerm** features are Boolean and can not be evaluated with AvP. We concluded that **SynTerm** feature is highly informative: if **SynTerm** (ab) = 1 then phrase ab is a domain term with probability of more than 80%.

We can see that in both cases the same set of features and use of machine learning methods lead to much higher values of average precision. However there are significant distinctions in ratios between AvP of features. For example, in

the banking domain AvP of the frequency feature has the highest value, features with high average precision in the science domain have enough low values in the banking domain.

We explain this phenomenon with relative narrowness of the banking domain. Banking documents contain a lot of terminology from neighbour domains such as economy or politics. So among extracted expressions, there are many real terms having all specific qualities of "termhood" but not related to the banking activity. In the scientific text collection the share of terms from other domains is much lower.

Also we can see failure of *SnipFreq_i* features in banking domain. The reason of this phenomenon, in our opinion, is as follows: the banking domain is subject to legal regulation, therefore documents of the domain contain a lot of citations from legal acts which leads to false large values of *SnipFreq_i*.

Table 1. Average Precision (AvP) for single features and logistic regression. Feature SnipTitle was not extracted for phrases in science domain.

Feature	AvP (Banking)%	AvP (Natural Sciences)%
Alphabet	40%	57%
Frequency	57%	66%
Mutual Information	43%	64%
MI_3	45%	67%
Inside	55%	75%
FreqBySnip	53%	69%
NearDefWords	49%	73%
Scalar ₁	42%	61%
Scalar ₂	45%	60%
BinarScalar ₁	49%	64%
BinarScalar ₂	48%	62%
Snipfreq ₀	34%	66%
Snipfreq ₁	38%	67%
Snipfreq ₂	38%	67%
Markers	40%	65%
Completeness	52%	69%
SnipTitle	50%	-
Logistic Regression	79% (+38.6% from Freq)	83% (+25.8% from Freq)

5 Conclusion

In this paper we have proposed to use three types of features for extraction of two-word terms and have shown that all these types of features are useful for term extraction. The set of features includes new features such as features extracted from the existing domain-specific thesauri and features based on Internet search results.

We have shown that the combination of several types of features considerably enhances the quality of the term extraction procedure. The developed system of term extraction reorders terms in a list of candidates much better than the basic-line order with decreasing frequency.

We studied the set of features for term extraction in two different domains. We found that developing term extraction models in a specific domain, it is important to take into account such properties of the domain as broad scope or narrow scope (science vs. banking) and relatedness to the socio-political domain, which is regulated with legal acts. We suppose that it is possible to find the main types of domains for term extraction, to select the best feature sets and corresponding machine learning models for every type of domains.

References

1. Zhang, Z., Iria, J.: Brewster, Ch., Ciravegna, F.: A Comparative Evaluation of Term Recognition Algorithms. In: Sixth International Language Resources and Evaluation, LREC 2008 (2008)
2. Pecina, P., Schlesinger, P.: Combining association measures for collocation extraction. In: Annual Meeting of the Association for Computational Linguistics, ACL 2006, ACM Press, New York (2006)
3. Dobrov, B., Loukachevitch, N.: Development of Linguistic Ontology on Natural Sciences and Technology. In: Linguistic resources and Evaluation conference, LREC 2006 (2006)
4. Manning, C., Raghavan, P., Shutze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
5. Daille, B., Gaussier, E., Lang, J.M.: An evaluation of statistics scores for word association. In: Tbilisi Symposium on Logic, Language and Computation, pp. 177–188. CSLI Publications (1998)
6. Nenadic, G., Ananiadou, S., McNaught, J.: Enhancing automatic term recognition through recognition of variation. In: 20th International Conference on Computational Linguistics (COLING 2004), pp. 604–610 (2004)